

Review of “A new proposal for optimizing maximum hydrological events fitting with flexible TCEV Distribution” by Ta, Valdes-Abellan, Yu and Li

Summary

The authors introduce a new estimation procedure for the parameters of the TCEV model. Their approach finds an initial estimate which is used to determine whether a TCEV should be applied. Once the decision has been made to fit a TCEV, a weighted RMSE is used to estimate the parameters and different weighting schemes are compared. Performance is assessed using 10 real-world data sets and some simulated data.

While the results show that the approach has potential benefits and the presentation is often clear, there are multiple aspects that require reworking. One of the arguments in Section 2.3.2 is not supported mathematically and several choices in the method appear arbitrary and require a more in-depth justification. It’s nice that the authors make the effort to analyse 10 different real-world data sets, but the uncertainty has not been quantified in much detail, which should be done for a full statistical analysis. The authors also comment on predictive performance, but this aspect should be best explored via a simulation study due to the limited size of the data set. In its current form, I don’t find the simulation study particularly interesting, and I would prefer a simulation study that focuses on a comprehensive comparison of the TCEV estimation methods considered in the 10 case studies. Finally, I would also like to see results discussed a bit more in context rather than just being focused on the authors’ own performance metrics. Please find my detailed comments below

Major Comments

- **Section 2.3.2:** The authors argue that Equation (3) implies that under a double-log transformation, the TCEV is described by two linear segments. However, this statement is not mathematically correct because this transformation gives:

$$-\ln[-\ln(F_{TCEV}(x))] = -\ln[\exp(-\alpha_1(x - \beta_1)) + \exp(-\alpha_2(x - \beta_2))],$$

and the right-hand side does not simplify to two linear segments. Therefore, while the argument in Equation (3) is correct, we require strong assumptions for the right-hand side to

approximately have the form of two linear segments. Specifically, we require the two Gumbel distributions to be sufficiently different. The authors need to carefully explore this aspect and provide a more rigorous analysis under which conditions this two-line segment representation can be assumed. Plotting some of the estimated models in Table 1, the necessary assumption does not seem to hold for all the applications. For instance, when I use the estimates from Case 1 in Table 1 using the equation above, I obtain a curve that looks close to linear, and it is noticeably different from the estimate in Figure 4 the authors obtain with their method. This aspect also results in follow-on issues related to Equations (4) and (5). Since the linear segments in Equation (4) are at best an approximation, it is not clear how the Gumbel parameters can be directly recovered. This probably requires some choice on how segments should be defined, but this aspect is not clearly explained. While I understand that these methods are mainly used to identify a starting point for the parameter values, this aspect is important, also because for the subsequent decision on whether a TCEV model should be considered.

- **Slope-Ratio:** The slope-ratio criterion of $R < 1.5$ feels very arbitrary to me, and the authors provide little justification for this choice. Should we not also consider the difference in β_1 and β_2 when deciding whether a simpler distribution may be used?
- **Weighting and model fitting:** The authors propose three weighting schemes, but it's not clear why these were selected in particular. Have the authors considered other schemes that just did not perform well in practice? A more serious issue for me is the overall fitting process. Some of the methods in Section 2.2 consider likelihood methods which allow us to account for the uncertainty in the observations, which tends to be very high for extreme observations. The approach by the authors considers the data as fixed as they look for the best match in terms of the empirical CDF. How is overfitting in such a situation avoided? I cannot find a discussion on this aspect in the manuscript.
- **Section 2.3.4** The authors select amongst the three weighting schemes by considering the fit across three evaluation areas, $[0.8, 1.0]$, $[0.6, 0.8]$ and $[0, 1]$ which again feels quite arbitrary. Have the authors explored sensitivity of their results in later sections to the choice of weighting given to these three areas?
- **Benchmark methods:** It is unclear how the POT method was applied. Have the others only applied it to the annual maximum series, or the full data set. I am also missing a discussion on which threshold selection techniques were used to determine the 80th percentile for the analysis. Can the authors please add this information to Subsection 2.4? Finally, would it not be useful to compare against the other TCEV fitting methods too, instead of just considering methods that cannot handle data generated from two different processes?
- **Simulation Setup:** In Section 3.2, the authors state that they also considered simulated data sets of size 30 years. Would one really consider the TCEV with four parameters for such a small

data set due to the large uncertainty in the parameter estimates? Regarding my comment above on the benchmark methods, I am also not sure why the authors sampled from a Pareto distribution which is usually used to model POT and not annual maxima.

- **Discussion on predictive performance:** At several places the authors state that their method has a better predictive performance. However, their assessment is purely based on data fitting. Assessing predictive performance in extreme value analysis is generally tricky. For the cases with a larger number of data points, one may consider assessing predictive performance using cross-validation although such an approach may be debated. In any case, it would be good to add confidence intervals in Figure 9 to assess uncertainty in the predictions obtained by the different methods.
- **Section 4.4:** The authors compare their method against other methods not involving the TCEV. However, it's not clear to this referee whether Figures 11 and 12 are an average over all the distributions described in Section 3.2, or only for a certain distribution. There is also a lack of detail when it comes to the parameter settings considered. Overall, I would have preferred a comparison of the different TCEV methods on simulated data, as this would have allowed the authors to properly assess predictive performance of the fitted models. Would the authors be able to run such a comparison, in which case the current Section 4.4 may be replaced.
- **Discussion of results:** The authors consider 10 Cases involving real-world data, but they do not discuss whether their conclusions agree with climatological knowledge. For instance, they argue that a TCEV is a suitable model for 8 out of 10 cases. I would like to see some comment on whether other studies also found that extremes at these sites are governed by two processes.

Minor Comments

- Line 41: The word “a” is missing before “Generalized Pareto Distribution (GPD)”.
- Line 106: It's more common to write “log-likelihood function” than “logarithmic maximum likelihood function”.
- Line 108: The text suggests that the second method is to maximize the “negative log-likelihood function”, but that does not align with the general statistical approach. If you were to use the negative log-likelihood function, you would try to minimize it. It also does not appear to me that methods 1 and 2 are really distinctly different, because they both consider the log-likelihood function.
- Line 144: I would suggest to write “where $\alpha > 0$ is the slope parameter and $(-\alpha\beta)$ is the intercept value of the line at the origin.”

- Table 1: Can the authors provide confidence intervals or standard errors for the standard errors please? This allows the reader to make a full comparison of the methods. Moreover, results for Cases 7 and 8 are also not presented, and no reason for their exclusion is provided, while this is done for Case 10.
- Figure 4: It seems that for some of the cases, only around 5 data points are used for Segment 2. How confident should we be in estimates for a Gumbel distribution based on only 5 data points given the high uncertainty in these observations? Further, the authors identified break points, but it's not clear how they are estimated in practice.
- Figure 6: The authors comment on the differences in performance for the different weighting schemes, but one may argue that the results in Figure 6 are in line with Figure 2 which shows that the quadratic weighting assigns the most weight to the most extreme observations. As such, the estimation will be pushed more towards matching these values in the quadratic weighting scheme than the other two. Maybe it would be good for the authors to refer the reader back to Figure 2 in their interpretation.
- Figure 7: It seems the fit for the highest values is not always great. Could the authors add confidence intervals to the quantile-quantile plot so that the reader can assess the model fit for rigorously?
- Section 4.3: The authors state that their method outperforms the other methods, but Figure 8 suggests that this does not apply for Cases 5 and 6 in terms of the RMSE. One may also argue that this form of performance comparison is skewed towards the proposed method, because a part of the estimation procedure considers three of the intervals for $F(x)$ on which the methods are compared when selecting the weighting method, and the performance in Figure 9 is exactly the one other one which optimizes.
- Figure 9: Please check that the sentence “The diamond symbols represent observed” in the caption is not missing it at least one word.
- Line 446: The authors state that the paper by Hüser and Wadsworth (2022) mentions weighting, but this paper considers a very different framework to the one in this paper. Please elaborate on the connection.
- The authors should please check their references. For instance, the Coles (2001) reference in Line 532 is incorrect.